Distributed Data Mining for Large NASA Databases

Hillol Kargupta
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County

Krishnamoorthy Sivakumar
School of Electrical Engineering and Computer Science
Washington State University

Outline

- Technical Objectives
- Problem Statement
- Technical Approach
- Data and NASA relevance
- Accomplishments and Preliminary Findings
- Technical Significance
- Web sites
- Facilities and Personnel Assigned
- Reference

Technical Objectives

- Facilitate knowledge discovery process from distributed large-scale earth science data using distributed data mining approach.
  - Development of new distributed data mining framework for massive heterogeneous distributed data.
  - Development of scalable distributed data mining algorithms.
    - Bayesian Networks (BN).
    - Decision Trees.
  - Development of techniques that translate the mined results to actionable knowledge.

Distributed Data Mining

- Motivation
  - Datasets are in distributed sites
  - Multiple Data streams: continuous source of data (e.g. over time) from multiple sources
  - Proactive data mining: A system that facilitates the exploratory nature of data mining, particularly using multiple data sources, by discovering possible inter-relationships between features across (distributed) datasets.
  - Example
  - NASA DAAC: NASA Distributed Active Archive Centers process, archive, document, and distribute data from NASA's Earth science research satellites and field measurement programs.

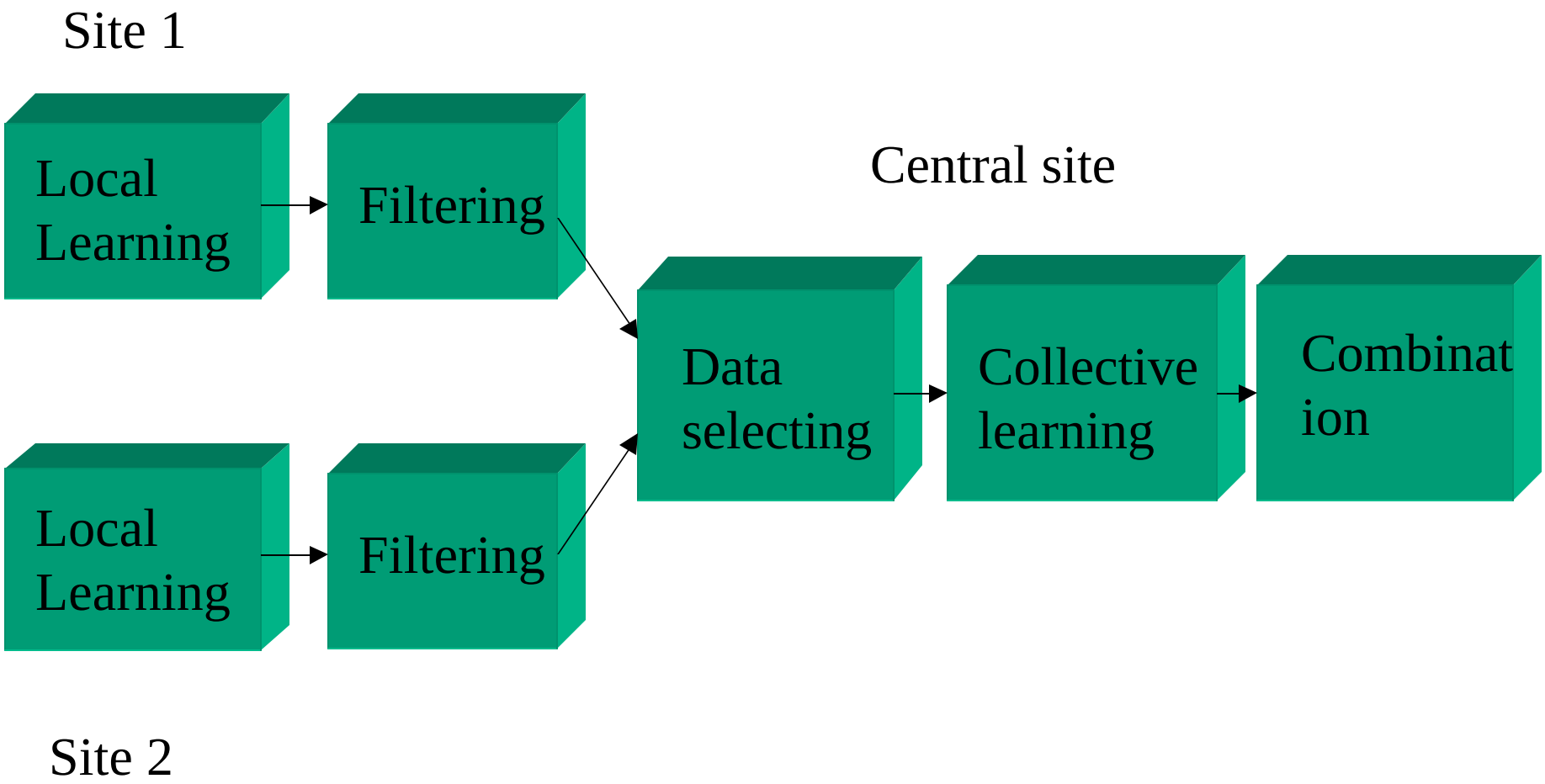
Problem Statement

- NASA DAACs provide a physically distributed but logically integrated database to support interdisciplinary research.
- Considering the size of data, availability of bandwidth, and security issues, it is not feasible to apply any off-the-shelf data mining system that require central aggregation of all distributed data.
- The proposed solution is to develop distributed techniques that exploit local data analysis.
  - Requires the minimum communication overhead.
  - Aggregation of local analyses guarantees provably correct results.

Technical Approach

- Collective Data Mining (CDM) framework.
  - Local learning: learn local model at each local site.
  - Data selecting: use local model to filter data and get a subset of data, then transmit these data to a central site.
  - Collective model learning: use the subset of data to learn a global model in the central site.
  - Combination: combine the local and global model to get a collective model.
- Models investigated under CDM.
  - Bayesian Networks.
  - Decision Trees.

Collective Data Mining



Data and NASA Relevance

- Two heterogeneous data sets are used for the initial investigation.
  - NASA DAO atmosphere data.
  - NOAA AVHRR Pathfinder data.

Data Description

- Data Assimilation Office (DAO).
  - Provides comprehensive and dynamically consistent datasets which represent the best estimates of the state of the atmosphere at that time.
  - NOAA AVHRR Pathfinder Data.
  - Statistics of aerosol optical thickness over the oceans.
  - Statistics of shortwave absorbed radiations and outgoing longwave radiations.

Features in the Data Sets

- DAO.
  - Features: Total of 26 features.
    - Surface prognostic products (3).
    - Upper air prognostic products (5).
    - One layer diagnostic products (18).
  - Temporal Coverage: March 1980 - November 1993.
- AVHRR Pathfinder.
  - Features: Total of 9.
    - Absorbed Solar Flux (2), Aerosol Optical Thickness (1).
    - Outgoing Longwave Radiation (4).
    - Total Fraction Cloud Coverage (2).
  - Temporal Coverage: Sep. 1981 – Dec. 2000.

Accomplishments and Preliminary Findings

- **Bayesian Network**: Development of a collective learning algorithm for BN.
- **Decision Tree**: Development of a collective learning technique and a model simplification technique using Fourier analysis.
- **Experimental results**: Preliminary investigation of earth science data using:
  - Clustering analysis
  - Correlation analysis

Distributed Bayesian Network Learning

- A Bayesian network (Belief network) is a probabilistic graph model.
- Defined as a pair (G, p), where G=(V, E) is a directed acyclic graph (DAG). V - node set, E - edge set. P - probabilities.
- Two problems: Structure and Parameter learning.

Benefits of Bayesian Network

- Problem decomposition: a compact model to represent a joint distribution
- A learned BN has clear interpretation: conditional independence and cause/effect
- Incorporation of prior knowledge
- Combine the prior knowledge and dataset

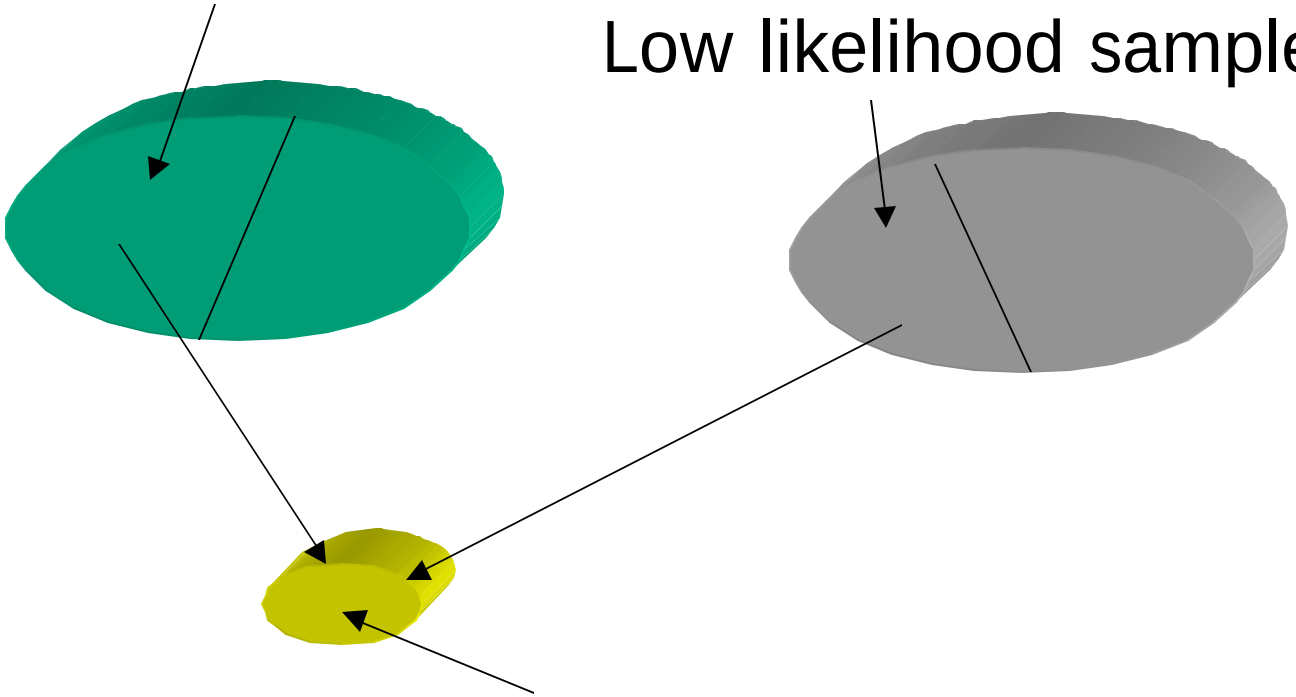
Collective Approach for BN Learning

- Compute local BN using local dataset
- At each site, identify the relatively low likelihood samples which is considered to be evidences of cross terms. Transmit part of these samples to central site
- Compute a non-local BN using these samples
- Combine local BN and non-local BN to get the collective BN

Data Filtering

Low likelihood samples in site A

Low likelihood samples in site B



Dataset transmitted to central site
(Intersection Set)

Key Points

- Low likelihood samples in local sites are the evidence of cross terms.
- Local dataset can learn local term well because of the conditional independence.

Implementation: Learning Software

- DistrBN: a distributed BN learning system based on C++ BN library SMILE.
- BODHI BN: working on incorporating BN learning to BODHI --- a distributed data mining system.

Decision Tree

- Decision tree is a famous predictive model.
- Decision tree can be used for large scale data efficiently.
  - Quickly constructible.
  - Fast to predict.
- Decision tree consists of a set of rules that can be easily translated into actionable knowledge.

Fourier Spectrum of Decision Trees

- A function can be represented as a linear combination of Fourier basis functions.
- A decision tree is essentially a function that is defined over discrete multivariate variable.
- Therefore a decision tree can be represented in terms of its Fourier spectrum.

Properties of Fourier Spectrum of Decision Trees

- Fourier spectrum of a decision tree is compact.
  - There exists polynomial number of non-zero Fourier coefficients.
  - A set of low-order Fourier coefficients is enough to represent the original decision tree.
- Fourier spectrum of a set of decision tree is the union of weighted spectrum of each tree.
- Therefore, multiple trees can be represented as a small Fourier spectrum.

CDM Approach To Construct a Distributed Classifier

- Construct local ensemble of decision trees at each site using Boosting.
- Together, all ensembles identify a set of data that should be centralized.
- An ensemble of decision tree is constructed from the centralized data.
- A prediction is made by collecting and combining all predictions from both local and central ensembles.

Pattern Discovery From Distributed Classifier

- We proved and implemented that a single tree can be constructed from a Fourier spectrum of multiple decision trees.
- Using the same approach, we can represent a distributed classifier with a single decision tree.
- The resulting tree is beneficial in extracting a set of significant patterns that govern the characteristics of the distributed data.

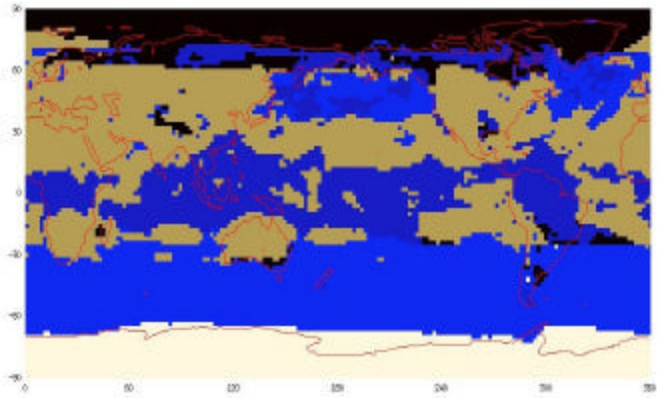
Preliminary findings

- Goal: Perform initial exploratory data analysis on real data using common off-the-shelf method.
- These preliminary findings would facilitate application of the developed CDM algorithms to real data.

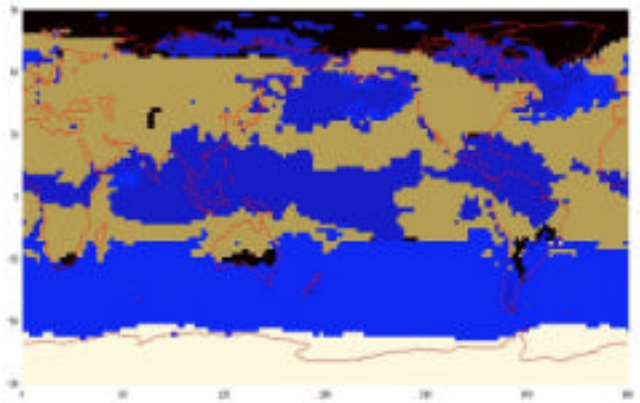
Clustering

- Goal:
  - Investigation of spatial correlation among grid points on the Earth.
- Input:
  - DAO and AVHRR Pathfinder data.
- Output:
  - Clusters of grids that exhibit similar behaviors.
- Algorithms used:
  - K-means.
  - EM, etc.

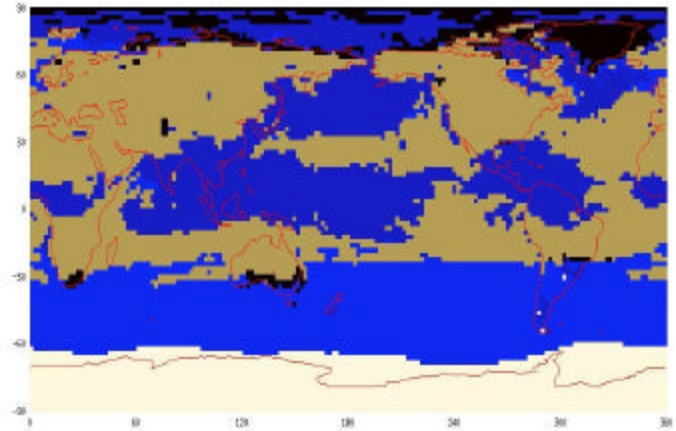
Clustering Result



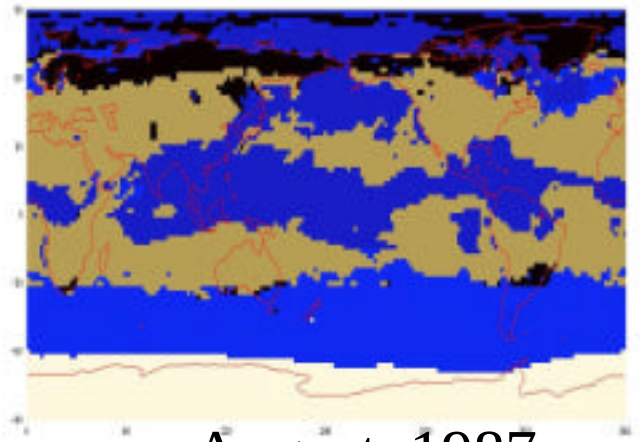
May, 1987



June, 1987



July, 1987



August, 1987

Clustering Example

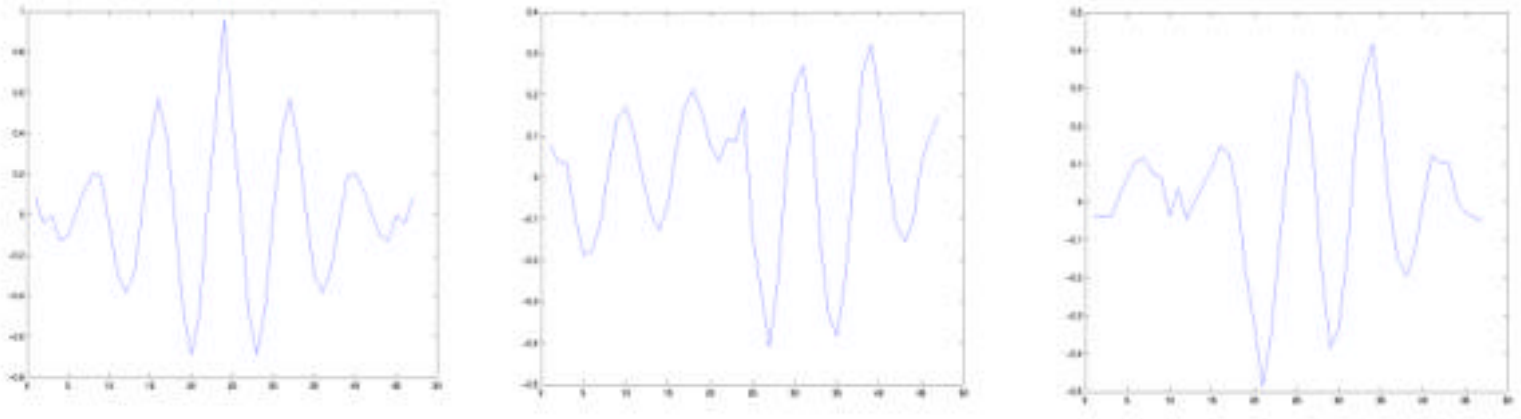
- During a fixed time period (Jan.-Dec., 1987), applied clustering algorithm to global DAO and NOAA.
- In our initial experiment, in May 87-Oct.87, a pattern (prominent cluster) was observed in the European (land) region and a region in Pacific ocean.
- These are very preliminary results and further investigation is underway.

Correlation Analysis

- Goal:
  - Detecting characteristics of earth science data by considering each feature as a random process.
- Auto-correlation analysis:
  - Detecting periodic behavior of each feature.
- Cross-correlation analysis:
  - Detecting correlation of a feature with other features.
  - This can be used to choose correlated sets of features for learning underlying models.

Correlation Analysis Example

- Use DAO data for time interval Jan.87- Dec.88
- Graphs depicted below (left to right): Feature cldfrc auto-correlation, Cross-correlation of feature cldfrc with tg, and that of feature cldfrc with vintvq



Correlation Analysis Example

- Auto-correlation graph indicates a strong periodic behavior.
- Cross-correlation graphs indicate a strong periodic correlation with lags.
- These are very preliminary results and further investigation is underway.

Technical Significance: BN

- A collective strategy to learn BN from distributed data.
- Association of features across data sites can be determined by local learning and selective transmission of data.
- The proposed approach scales well with respect to number of sites, features, observations (Chen et. al. 2001).
- This approach is well suited to a data stream scenario (Chen et. al. 2001a), where data is continuously updated.

Technical Significance: Decision Tree

- Decision tree construction using Fourier analysis.
  - Decision tree construction from ensemble-based distributed classifier.
  - Decision tree construction from distributed data.
  - Decision tree construction from data streams.
- The resulting tree enhances our ability to understand patterns contained in distributed data set.

Decision Tree Construction From Distributed Classifier

- The proposed decision tree-based distributed classifier adopts ensemble approach.
- It has been proven that ensemble of decision tree can be aggregated in the Fourier domain.
- The technique that constructs a single decision tree from aggregated spectrum of ensemble has been implemented [Park and Kargupta, 2002].

Decision Tree Construction From Data Streams

- Ensemble approach is a natural candidate to learn a classifier from data streams.
- Using the spectrum of ensemble, a intuitive and rich representation of data stream is possible [Kargupta and Park, 2001].
- From the ensemble that is constructed from a data stream, a single tree can be built using the same technique mentioned in the previous slide.

Decision Tree Construction From Distributed Data

- By computing Fourier spectrum directly from data, a decision tree can also be constructed.
- An algorithm that approximates Fourier spectrum directly from data has been implemented [Ayyagari and Kargupta, 2002].
  - Probabilistic approach to estimate Fourier coefficients from data with unknown distribution.
  - Maximize uniformity of underlying data through adaptive sampling.

Expected Impact On NASA

- This research would substantially enhance our ability to evaluate, analyze, and understand vast quantities of heterogeneous data that are archived in diverse, highly specialized, and geographically dispersed centers, into knowledge.
- It would generate theories to distill knowledge and information contained in such raw data, by detecting regularities, patterns, and correlations, identifying potential causal interactions, and extracting relevant features.
- This would greatly influence the understanding of vast amounts of data gathered by various NASA missions.
- It would also help in combining data from NASA sources and other related sources like NOAA and CDC, to understand the interrelationship between different features over land, ocean, and their impact on, for example, disease emergence.

Web sites

- http://www.cs.umbc.edu/~hillol/ddm.html
- http://www.eecs.wsu.edu/~hillol/diadic.html
- http://www.cs.umbc.edu/~hillol/diadic.html

Personnel Assigned

UMBC

- Dr. Hillol Kargupta (PI)
- Dr. Byung-Hoon Park (post doctoral research associate)

WSU

- Dr. K. Sivakumar (Co-PI)
- Rong Chen (graduate student)
- Jianjie Ma (graduate student)
- Meng Da (graduate student)

Facilities Used

- Computing facilities of the DIADIC (DIstributed Adaptive DIscovery and Computation) laboratories located in:
  - Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County (UMBC).
  - School of Electrical Engineering and Computer Science, Washington State University (WSU).
- Other basic infrastructure provided by UMBC and WSU.

References

- [Chen et.al, 2001] Chen, R., Sivakumar, K., and Kargupta, H. Distributed Web Mining using Bayesian Networks from Multiple Data Streams. Proceedings of the IEEE International Conference on Data Mining, 75--82. IEEE Press.
- [Chen et. al, 2001a] Chen, R., Sivakumar, K., and Kargupta, H. An Approach to Online Bayesian Learning from Multiple Data Streams, Proceedings of the Workshop on Mobile and Distributed Data Mining, PKDD2001.
- [Kargupta and Park, 2001] Kargupta, H. and Park, B. Mining Decision Trees from Data Streams in a Mobile Environment. Proceedings of the IEEE International Conference on Data Mining, 281--288. IEEE Press.

References

- [Park and Kargupta, 2002] Park, B. and Kargupta, H. Constructing Simpler Decision Trees from Ensemble Models Using Fourier Analysis, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).
- [Ayyagari and Kargupta, 2002] Ayyagari, R. and Kargupta, H. A Resampling Technique for Learning the Fourier Spectrum of Skewed Data, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).